Article

*Discovery Medicine* 2025; 37(199): 1548–1558
https://doi.org/10.24976/Discov.Med.202537199.134

# Translating Proteomic Profiles Into Therapeutic Targets: A Mendelian Randomization Analysis in Breast Cancer

Qian Min[1,†], Yi Dong[1,†], Qingqing Li[1,*], Yi Yao[1,*]

[1]Cancer Center, Renmin Hospital of Wuhan University, 430060 Wuhan, Hubei, China
*Correspondence: qqli1130@whu.edu.cn (Qingqing Li); yaoyi2018@whu.edu.cn (Yi Yao)
[†]These authors contributed equally.
Published: 20 August 2025

**Background: Breast cancer remains one of the most commonly diagnosed malignancies, yet its underlying mechanisms are incompletely understood. This study aimed to investigate the causal relationships between circulating proteins and breast cancer using Mendelian Randomization (MR) and to identify potential therapeutic targets.**
**Methods: Genomic data for breast cancer from the FinnGen study and protein quantitative trait loci (pQTL) data from Decode were analyzed. MR was applied to evaluate the effects of specific proteins on breast cancer risk, followed by validation through colocalization analyses and summary-data-based Mendelian Randomization (SMR). Two-step MR was used to assess the mediating role of immune cells in the causal pathway between the target proteins and breast cancer. Moreover, expression patterns and prognostic significance of the identified proteins in breast cancer were examined.**
**Results: Several proteins [glutathione S-transferase mu (GSTM) 1, N-acylsphingosine amidohydrolase 2 (ASAH2), TNF superfamily member 12 (TNFSF12), hematopoietic prostaglandin D synthase (HPGDS), nidogen 2 (NID2), SEMA4D, snurportin 1 (SNUPN), GSTM3, MYC-associated factor X (MAX), pregnancy zone protein (PZP), GSTM4, and uroporphyrinogen decarboxylase (UROD)] were significantly associated with breast cancer. Colocalization and SMR analyses supported the potential role of GSTM in breast cancer, especially GSTM1 ($p_{SMR}$ = 0.02 and $p_{HEIDI}$ = 0.08). Mediation MR implicated CD11c-expressing monocytes ($Z$ = –2.197, $p$ = 0.028) in the GSTM-breast cancer pathway. Notably, lower GSTM1/4 expression levels correlated with poorer clinical outcomes in breast cancer patients.**
**Conclusion: This study used cis-pQTL MR and colocalization to identify GSTM1/4 as a potential druggable gene and promising therapeutic target for breast cancer, which might be affected by immunity. These findings contributed to understanding the role of metabolic and immune mechanisms in breast cancer.**

**Keywords:** breast cancer; protein; GSTM; Mendelian Randomization; therapy

## Introduction

Epidemiological evidence indicates that breast cancer persists as one of the most commonly diagnosed malignancies and remains the leading cause of cancer-related mortality among women globally [1]. Despite advancements in survival, outcomes still need to be improved, influenced by heterogeneous clinical factors, biological diversity, and variable therapeutic responses [2]. Current treatment is primarily based on a comprehensive assessment of tumor characteristics, including T stage, metastatic status, and biomarkers such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). However, significant challenges, including therapeutic resistance, remain unresolved, primarily undermining clinical efficacy, increasing tumor recurrence and metastasis, and ultimately raising mortality [3]. With innovations in experimental techniques, investigating molecular-level mechanisms and identifying predictive biological biomarkers alongside targeted therapies have become promising directions to optimize therapeutic strategies and improve long-term outcomes.

All individuals possess a range of genetic variations that are nearly completely randomized, some of which influence biological traits. By applying statistical approaches such as Mendelian Randomization (MR) to trace the effects of instrumental variables, it is possible to deduce the generally complex relationships between the explanatory and dependent variables [4]. The combined application of genome-wide association studies (GWAS) and MR offers a novel strategy for screening potential therapeutic targets for breast cancer [5]. MR, as a causal inference method leveraging genetic variants as instrumental variables, elucidates etiological relationships between exposures and outcomes, thereby facilitating the translation of epidemiological findings into precision oncology [6]. Through the integration of statistical methodologies, including causal inference and colocalization analyses, the underlying mechanisms of breast cancer can be clarified [7]. In this study, we combined protein quantitative trait loci (pQTL) data

Copyright: © 2025 The Author(s). Published by Discovery Medicine. This is an open access article under the CC BY 4.0 license.
Note: *Discovery Medicine* stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

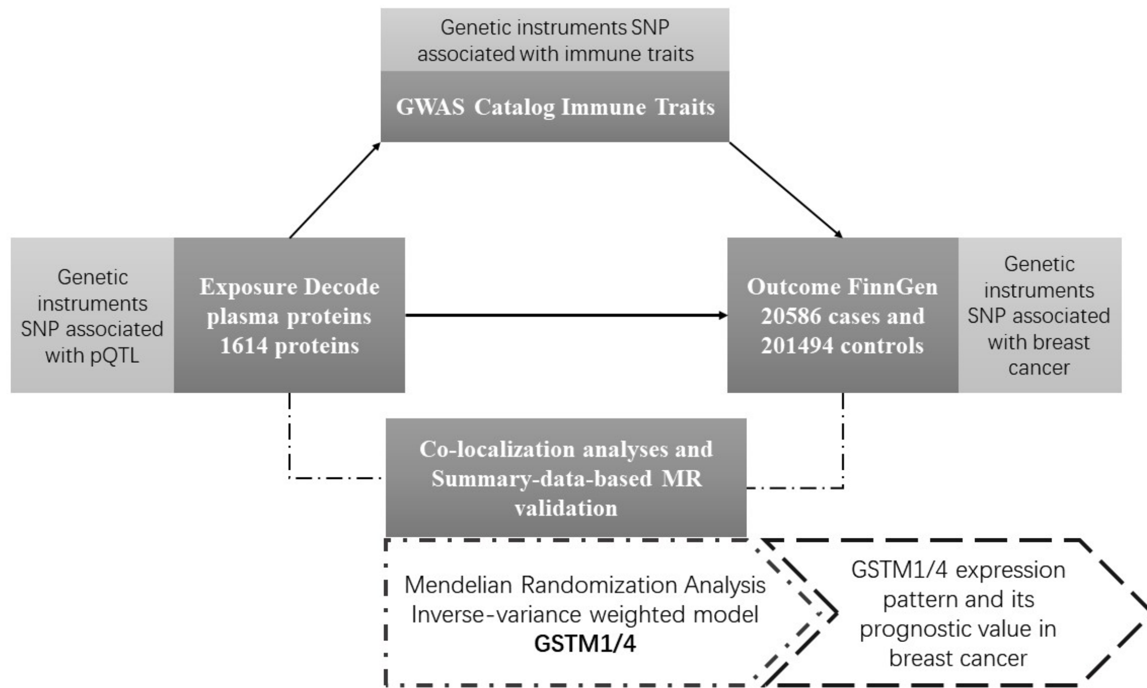


**Fig. 1. Flowchart illustrating the proteome-wide identification of novel therapeutic target proteins for breast cancer using Mendelian Randomization analysis.** SNP, single-nucleotide polymorphism; pQTL, protein quantitative trait loci; GSTM, glutathione S-transferase mu.

with the FinnGen database to identify target proteins with potential applications in drug development for breast cancer. Moreover, we employed a comprehensive analytical framework, including colocalization, summary-data-based Mendelian Randomization (SMR), and mediation MR, to demonstrate the causal relationship of GSTM and breast cancer, as well as its prognostic value. Moreover, it underscored the pivotal role of metabolic-immune crosstalk in breast cancer pathogenesis, highlighting its significance for future therapeutic exploration.

## Methods

First, we downloaded pQTL data and breast cancer data and then filtered cis-pQTLs for further analysis. Next, we performed MR analysis based on these cis-pQTLs to screen target proteins. Colocalization analysis and SMR were applied for further validation. Moreover, this study employed mediation MR to systematically explore potential mechanisms of breast cancer based on publicly accessible proteomic and genomic databases, as well as the role of immune cells. The multi-stage investigation workflow, encompassing genetic instrumental variable selection, colocalization analysis, and clinical value exploration, is presented in Fig. 1. All datasets used in this study had undergone formal institutional review board approvals and obtained informed participant consent in their original study contexts.

### *Data Collection and Instrument Selection for MR*

The pQTL data were sourced from the Decode (https://www.decode.com/summarydata/) [8]. Genotype data were obtained from FinnGen to provide novel insights into the genetics of breast cancer. Preliminary screening focused on breast cancer-associated proteins, with breast cancer itself as the outcome rather than molecular subtypes. GWAS summary statistics for breast cancer were obtained from the FinnGen Research Project (R11), comprising 20,586 patients and 201,494 controls (https://storage.googleapis.com/finngen-public-data-r11/summary_stats/finngen_R11_C3_BREAST_EXALLC.gz). GWAS data of immune cell traits (Ebi-a-GCST0001391 to Ebi-a-GCST0002121), encompassing 731 traits, were retrieved from the publicly available GWAS Catalog (https://gwas.mrcieu.ac.uk/) [9]. This dataset captured the effects of 22 million genetic variants in 3757 Sardinian individuals and included several immune cell types: T cells, B cells, dendritic cells, and myeloid cells, among others. Quantitative analysis of these 731 traits revealed distinct measurement categories applicable to pharmaceutical research and expanded the spectrum of immune cell-related features, as well as the cellular relationship between immunity and the risk of disease.

Using the "TxDb.Hsapiens.UCSC.hg38.knownGene" package, our selection criteria prioritized cis-acting pQTL variants (located within $\pm$1000 kb of transcription start

DM DISCOVERY MEDICINE 2001

sites) to enhance biological feasibility. Variants meeting the following screening criteria were chosen as instrumental variables (IVs). First, we incorporated single-nucleotide polymorphisms (SNPs) that achieved genome-wide significance ($p < 5 \times 10^{-8}$). When no sufficiently significant SNPs were available across the genome that could serve as IVs, slightly less stringent thresholds ($p < 1 \times 10^{-6}$) were considered. Second, the F-statistic, which quantifies the joint ability of genetic instruments to explain exposure variance, was set at a minimum threshold of 10 to reduce bias from weak instruments. Third, linkage disequilibrium (LD), representing the non-random association between different SNPs and potential confounders in MR analyses, was accounted for using a clumping approach ($r^2 < 0.01$ and clumping distance = 10,000 kb). Harmonization was performed using the "Two Sample MR" R package to align allelic effect directions of SNP instruments between the exposure and outcome datasets.

### *Mendelian Randomization and Validation*

We aimed to identify the potential mechanisms of the selected proteins in breast cancer risk, focusing on results derived from the Inverse Variance Weighted (IVW) method as the primary analytical approach. Bidirectional MR analyses were first conducted to explore the causal relationship between the proteins and breast cancer. Unless otherwise stated, the MR results reported in this study were based on IVW. MR findings were presented as beta estimates with standard errors and odds ratios (OR) with 95% confidence intervals (CI). However, since we selected cis-pQTLs, there were insufficient SNPs for the exposure in reverse analysis. To further assess underlying mechanisms, we employed a two-step MR to perform mediation analysis, aiming to evaluate the potential role of immune cells as intermediaries in the causal pathway connecting the target protein to breast cancer risk. The effects of each target protein on breast cancer were divided into two components: direct effects caused by the target protein and indirect effects mediated through immune cells. First, we computed the causal effect of each target protein on immune cells ($\beta 1$), followed by the causal influence of immune cells on breast cancer ($\beta 2$). The mediation effect was calculated as $\beta 1 \times \beta 2$, and its proportion relative to the total effect was expressed as (mediation effect / total effect) $\times$ 100%, with 95% CIs provided.

MR analyses were conducted using the R packages of "Two Sample MR" and "Mendelian Randomization" in R software (version 4.3.1, R Foundation, Vienna, Austria). MR-Egger regression was used to assess and adjust for bias caused by heterogeneity. Heterogeneity was evaluated using Cochrane's Q-test, with smaller $p$-values indicating greater heterogeneity and a higher likelihood of directional pleiotropy. Leave-one-out analyses were conducted to identify potential SNP outliers. Ambiguous or duplicate SNPs were removed, and alleles were mapped to the human genome reference sequence (build 37).

### *MR Validation Based on Colocalization Analysis and SMR*

Colocalization and SMR analyses were used to assess causal relationships between genes and breast cancer. Firstly, concordance between protein levels and breast cancer risk was evaluated through colocalization analysis to ensure that observed associations were not confounded by underlying genetic factors. A Bayesian test was performed on the preliminary results (selected proteins) using the "coloc" R package. To avoid unstable estimates due to rare variants, SNPs with minor allele frequency (MAF) <0.05 were excluded. For colocalization analysis, we used SNPs associated with pQTL and breast cancer ($p < 5 \times 10^{-6}$). The Posterior Probability of Haplotype (PPH) was a key metric in colocalization analysis, representing the posterior probability that two traits share a causal variant. PPH was categorized into five classes: PPH0: no association with either trait; PPH1: associated only with protein expression; PPH2: associated only with disease risk; PPH3: associated with both traits but at different loci; PPH4: associated with both traits sharing the same causal variant. Commonly applied thresholds included PPH3 + PPH4 >0.7, indicating a high probability of colocalization, and PPH4 >0.5, indicating a high probability of a shared causal variant. Given the limited data from cis-pQTLs, a more lenient threshold was applied, and variants with PPH3 + PPH4 >0.5 or PPH4 >0.2 were selected for further investigation.

The causal relationship between potential protein expression and breast cancer was further validated using the online SMR software tool (https://yanglab.westlake.edu.cn/software/smr/#DataResource) [10]. To improve the reliability, the Heterogeneity in Dependent Instruments (HEIDI) test was applied. Associations with $p_{HEIDI} < 0.05$ suggested potential pleiotropy and were therefore excluded.

### *Single-Cell Analysis and Clinical Value*

Single-cell transcriptomic data were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205472), specifically dataset GSE205472. The sequencing data of 13 breast cancer tissue samples (GSM6213322–GSM6213334) were integrated and processed using the "Seurat" R package. After normalization with the LogNormalize method and correction for batch effects, the cells were clustered into 25 subpopulations and annotated based on known marker genes. Clusters were visualized using T-distributed stochastic neighbor embedding (T-SNE) as implemented in Seurat. The FindAllMarkers function was applied to compare gene expression levels between cell populations, performing differential expression analysis for each subpopulation. A gene was considered expressed if present in at least 25% of cells. Marker genes were defined by an average $Log_2FC$ >0.5 and a false discovery rate (FDR)-adjusted $p$-value $< 0.05$.

To evaluate gene expression and clinical significance in breast cancer (BRCA), we used the UALCAN web por-

tal (http://ualcan.path.uab.edu), a comprehensive resource for exploring The Cancer Genome Atlas (TCGA) transcriptomic and clinical datasets [11]. The Promoter Methylation module was also utilized, analyzing average $\beta$ values from the Illumina Infinium HumanMethylation450 (HM450) platform, where $\beta = 0$ indicates unmethylated and $\beta = 1$ indicates fully methylated states for CpG islands in promoter regions of selected genes.

For survival analysis, breast cancer patients were divided into high- and low-expression groups based on the median expression value of the target gene, and Kaplan-Meier survival analyses were performed using the KM-Plotter platform (https://kmplot.com/analysis/) [12].

To assess therapeutic potential, target genes were submitted to the Drug Signatures Database (DSigDB, http://dsigdb.tanlab.org/DSigDBv1.0/) for evaluating protein-drug interactions [13]. This evaluation of the interactions between pharmaceuticals and proteins was essential in determining whether the identified proteins could be effectively applied. Additionally, the genes were uploaded to the Enrichr platform (https://maayanlab.cloud/modEnrichr/) to predict potential drugs that may target these proteins [14].

## Results

### *Proteomic Analysis and Identification of Breast Cancer-Related Proteins and Mediated MR*

We identified 4706 proteins from cis-pQTL data. After eliminating linkage disequilibrium, 1614 proteins met the criteria for further analysis. Using rigorous statistical methods, including the IVW approach and Bonferroni correction, our study identified significant associations between breast cancer risk and 13 plasma proteins ($p < 3.1 \times 10^{-5}$, 0.05/1614), as shown in Table 1 and **Supplementary Table 1**. These proteins included pregnancy zone protein (PZP), glutathione S-transferase mu (GSTM) 3, semaphorin 4D (SEMA4D), proenkephalin (PENK), N-acylsphingosine amidohydrolase 2 (ASAH2), TNF superfamily member 12 (TNFSF12), MYC-associated factor X (MAX), hematopoietic prostaglandin D synthase (HPGDS), GSTM1, nidogen 2 (NID2), snurportin 1 (SNUPN), GSTM4, and uroporphyrinogen decarboxylase (UROD).

Among these, higher levels of PZP, GSTM3, GSTM1, SNUPN, and GSTM4 were associated with a reduced risk of breast cancer, whereas elevated levels of SEMA4D, PENK, ASAH2, TNFSF12, MAX, HPGDS, NID2, and UROD were associated with an increased risk. These findings were consistent across weighted mode, weighted median, and MR-Egger analyses. Except for PENK, the $p_{\text{pleiotropy}}$ of all other proteins was >0.05. However, heterogeneity analysis revealed $p$-values $< 0.05$ for GSTM3, SEMA4D, ASAH2, HPGDS and GSTM4 (**Supplementary Table 2**). We speculated that such heterogeneity might be due to inconsistent effects among IVs (e.g., population stratification or differences in breast cancer subtypes). Due to the limited number of protein-associated SNPs, reverse MR analysis did not yield significant results.

**Table 1. Mendelian Randomization (MR) results for proteins significantly associated with breast cancer.**

| Protein | OR (95% CI) | $p$-value (IVW) | Mean F-statistic |
|---|---|---|---|
| PZP | 0.941 (0.922–0.959) | $7.27 \times 10^{-10}$ | 207.668 |
| GSTM3 | 0.936 (0.917–0.957) | $1.41 \times 10^{-9}$ | 155.700 |
| SEMA4D | 1.065 (1.042–1.089) | $2.29 \times 10^{-8}$ | 316.588 |
| PENK | 1.058 (1.036–1.08) | $1.32 \times 10^{-7}$ | 217.975 |
| GSTM1 | 0.954 (0.936–0.971) | $3.14 \times 10^{-7}$ | 200.089 |
| NID2 | 1.056 (1.035–1.078) | $1.99 \times 10^{-7}$ | 184.011 |
| ASAH2 | 1.059 (1.036–1.082) | $2.80 \times 10^{-7}$ | 276.406 |
| TNFSF12 | 1.147 (1.086–1.212) | $9.95 \times 10^{-7}$ | 182.053 |
| UROD | 1.146 (1.082–1.213) | $2.98 \times 10^{-6}$ | 108.871 |
| SNUPN | 0.791 (0.715–0.875) | $4.83 \times 10^{-6}$ | 73.432 |
| MAX | 1.203 (1.107–1.308) | $1.46 \times 10^{-5}$ | 82.173 |
| HPGDS | 1.057 (1.031–1.084) | $1.75 \times 10^{-5}$ | 317.017 |
| GSTM4 | 0.960 (0.941–0.978) | $2.77 \times 10^{-5}$ | 191.560 |

OR, odds ratio; CI, confidence interval; PZP, pregnancy zone protein; GSTM, glutathione S-transferase mu; SEMA4D, semaphorin 4D; PENK, proenkephalin; ASAH2, N-acylsphingosine amidohydrolase 2; TNFSF12, TNF superfamily member 12; MAX, MYC-associated factor X; HPGDS, hematopoietic prostaglandin D synthase; NID2, nidogen 2; SNUPN, snurportin 1; UROD, uroporphyrinogen decarboxylase.

We further applied multivariable MR (MVMR) to identify immune cells that could independently influence breast cancer and computed the mediated effects and proportions of immune cells in the causal pathway between selected proteins and breast cancer. After screening, "CD11c on monocyte" was found to mediate the association between GSTM1 and breast cancer, with a mediation proportion of 9.5% ($p = 0.028$), as well as between GSTM4 and breast cancer (mediating effect: 12.1%, $p = 0.026$) (Table 2). No significant immune mediators were identified for other proteins affecting breast cancer, including HPGDS, MAX, and TNFSF12.

### *Confirmation of Therapeutic Protein Targets by Colocalization Analysis and SMR Analysis*

To validate the role of GSTM in breast cancer, we applied colocalization analysis and SMR for further analysis. Colocalization analysis linked multiple SNPs to two different traits, while data filtering was used to remove low-quality SNPs that might influence outcomes. The $p$-value threshold for SNPs associated with proteins and breast cancer was set at $<5 \times 10^{-6}$. Based on colocalization analysis, we confirmed shared genetic variants associated with breast cancer risk (Fig. 2).

Among the 13 potential causal proteins identified through genome-wide proteome MR analysis, three proteins (GSTM3, PZP and UROD) were excluded from colo-

DM DISCOVERY MEDICINE 2001

**Table 2. Mediated Mendelian Randomization (MR) assessing immune cell mediation between GSTM expression and breast cancer.**

| cis-pQTL | Mediator | nSNP | $\beta$_all | $\beta 1$ | $\beta 2$ | $\beta$_dir | SE | Z | *p*-value |
|---|---|---|---|---|---|---|---|---|---|
| GSTM1 | CD11c on monocyte | 138 | –0.0476 | –0.0699 | 0.0693 | –0.0427 | 0.0022 | –2.197 | 0.028 |
| GSTM4 | CD11c on monocyte | 137 | –0.0413 | –0.0723 | 0.0693 | –0.0363 | 0.0022 | –2.229 | 0.026 |

SNP, single-nucleotide polymorphism.

calization analysis due to the absence of overlapping SNPs with MAF >0.05 and inability to construct an LD matrix for the target region. In most cases, the colocalization posterior probability (PP.H4.abf) was below 0.8, reaching only 0.341 in HPGDS, thus failing to reach statistical significance. This suggested a possible but inconclusive likelihood of shared causal variants between proteins and breast cancer. However, unidirectional association evidence (PP.H1.abf) was relatively strong in most cases, typically exceeding 0.7, with the highest reaching 0.95 (TNFSF12), indicating a high likelihood of protein influence on traits, implying they may directly affect breast cancer through specific mechanisms. Additionally, independent association evidence (PP.H3.abf) demonstrated a modest degree of probability in some proteins, such as HPGDS and GSTM4, reaching 0.23 and 0.25, respectively. Notably, GSTM1, HPGDS, and ASAH2 exhibited the highest PPH4 values, indicating a higher probability of shared causal variant between protein levels and breast cancer risk (**Supplementary Table 3**). The no-association evidence (PP.H0.abf) was near zero in all results, further supporting a potential association between genes and breast cancer.

Overall, while colocalization evidence was insufficient to establish strong causality, unidirectional association evidence supports the hypothesis that these genes may influence breast cancer. Given our focus on cis-pQTL, we further applied SMR for validation. SMR results indicated that several SNPs showed statistically significant associations with gene expression, though further specificity validation is required. For example, SNP rs145858840 of GSTM1 showed a significant association with gene expression but nonsignificant colocalization ($p_{SMR}$ = 0.02 and $p_{HEIDI}$ = 0.08, **Supplementary Table 4**), suggesting an independent association. Similarly, rs11101980 significantly influenced GSTM4 expression ($p_{HEIDI}$ = 0.064, $p_{SMR}$ = 0.124), indicating a possible colocalization gap that warrants further investigation (**Supplementary Table 4**). In summary, our findings suggest that GSTM acts as a protective factor in the progression of breast cancer. Therapeutic strategies targeting GSTM may offer a novel and effective approach to reducing breast cancer risk.

### *Investigation of Potential Mechanisms*

The two-step MR analysis investigated the causal relationship and potential mechanisms linking GSTM to breast cancer. Volcano plots showing the phenotypic effects of target proteins, including GSTM1/3/4, ASAH2 and HPGDS, are presented in Fig. 3A. Based on immune cell associations with breast cancer, CD11c expression showed a positive effect on monocytes in relation to breast cancer. The results indicated that CD11c on monocytes served as an intermediary in the causal pathway between GSTM1/4 and breast cancer ($p$ = 0.028, $p$ = 0.026, respectively).

We further investigated the cell-type-specific expression of GSTM-coding genes in breast cancer tissues using single-cell transcriptomics. Analysis of the GSE205472 scRNA-seq dataset, following rigorous quality control and normalization, identified key cell populations within the breast cancer tumor microenvironment (TME) using principal component analysis (PCA) for dimensionality reduction and Louvain clustering, annotated according to the Human Primary Cell Atlas (HPCA). Major populations included tumor epithelial cells, fibroblasts, T cells, macrophages, monocytes, B cells, NK cells, tissue stem cells, and endothelial cells, exhibiting distinct spatial heterogeneity in t-SNE visualization (Fig. 3B).

Notably, analysis of GSTM1 and GSTM4 expression across cell populations revealed distinct expression patterns. Although GSTM1 expression was relatively low (Fig. 3C), both GSTM1 and GSTM4 were predominantly expressed in epithelial cells, aligning with their established roles in detoxifying electrophilic compounds. Immune cells, particularly myeloid populations such as macrophages, demonstrated lower expression levels of GSTM1 and GSTM4 compared to epithelial cells (Fig. 3C,D).

### *Identification of GSTM1 Expression and Clinical Value in Breast Cancer*

The clinical significance of GSTM requires further investigation. Differential expression analysis compared GSTM1/GSTM4 mRNA expression [RNA-seq data, log-transformed Transcripts Per Million (TPM) values] between tumor samples (n = 1097) and adjacent normal breast tissue samples (n = 114) from TCGA BRCA. Statistical significance was assessed using UALCAN's built-in Student's *t*-test (Fig. 4A,B). Methylation $\beta$ values were compared between primary tumors (n = 793) and normal samples (n = 97) (Fig. 4C,D). Transcriptional downregulation

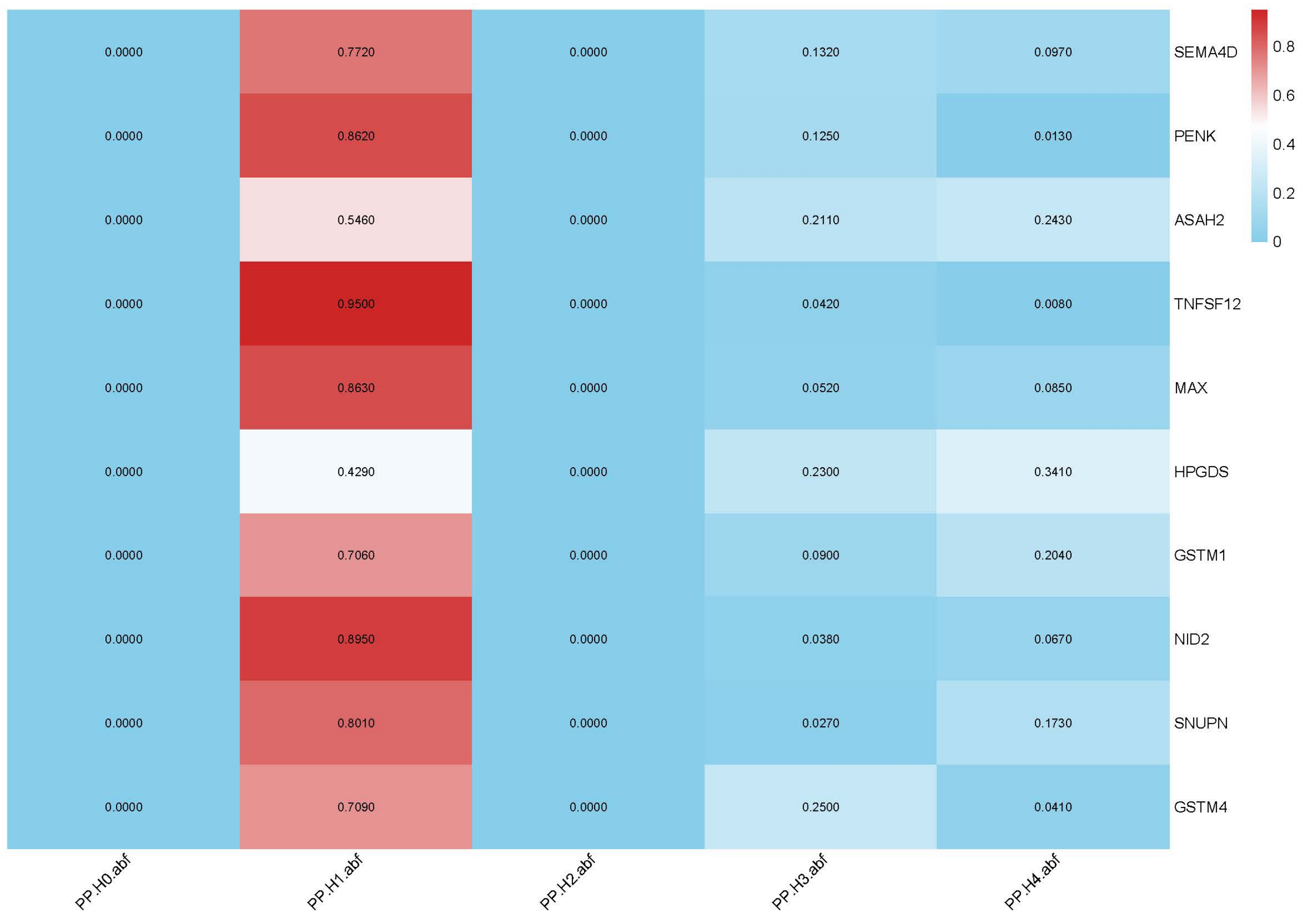


**Fig. 2. Colocalization analysis of gene-breast cancer associations.** Posterior probabilities for five hypotheses: H0, no association between the gene and breast cancer; H1, the gene influences the trait; H2, the trait influences the gene; H3, the gene and the trait are independently associated; H4, the gene and the trait colocalize, sharing a causal variant. A posterior probability (PP) value closer to 1 indicates stronger support for the corresponding hypothesis.

of GSTM4 and promoter hypermethylation of GSTM1 and GSTM4 were significantly associated with breast cancer. Analysis of TCGA breast cancer data revealed a marked reduction in GSTM4 mRNA expression in BRCA compared to normal tissue ($p < 0.001$), with distinct expression patterns across molecular subtypes. GSTM1 expression was significant lower in triple-negative breast cancer (TNBC) compared to normal and luminal subtypes ($p < 0.001$) (Fig. 4E). For GSTM4, no statistically significant difference was observed between normal and luminal groups, though average expression in the normal group was relatively higher (Fig. 4F). This transcriptional suppression was strongly linked to promoter hypermethylation, with significantly elevated $\beta$ values in tumor tissues compared to normal (GSTM1: $\beta_{tumor} = 0.53$ vs. $\beta_{normal} = 0.21$, $p < 0.001$; GSTM4: $\beta_{tumor} = 0.38$ vs. $\beta_{normal} = 0.37$, $p = 0.03$). Notably, low expression of either gene predicted significantly worse overall survival (Fig. 4G,J, GSTM1: HR = 0.56, 95% CI: 0.44–0.7, $p < 0.001$; GSTM4: HR = 0.53, 95% CI: 0.43–0.67, $p < 0.001$). Subgroup analysis indicated the association between low-expression of GSTM1/4 and poor prognosis in Luminal-subtype, highlighting its protective value (Fig. 4H,K). Conversely, within the TNBC, a statistically significant association was observed in GSTM1 ($p = 0.0014$) (Fig. 4I,K,L). These findings strongly implicate epigenetic silencing of GSTM1 and GSTM4 as a key mechanism contributing to breast cancer pathogenesis and progression, supporting their potential as biomarkers for aggressive disease and unfavorable patient outcomes. To evaluate the GSTM1 as a potential therapeutic target, drug-gene interaction analysis using DSigDB drug database identified several associated compounds, including suloctidil (CTD 00006994), Tamibarotene (CTD 00002527), and Decitabine (CTD 00000750) (Table 3).

## Discussion

Despite recent advances in anti-cancer therapies, breast cancer continues to significantly impair the quality of life for patients. Current treatments remain relatively ineffective, and deeper mechanistic insights are still lacking. Therefore, identifying novel therapeutic targets for

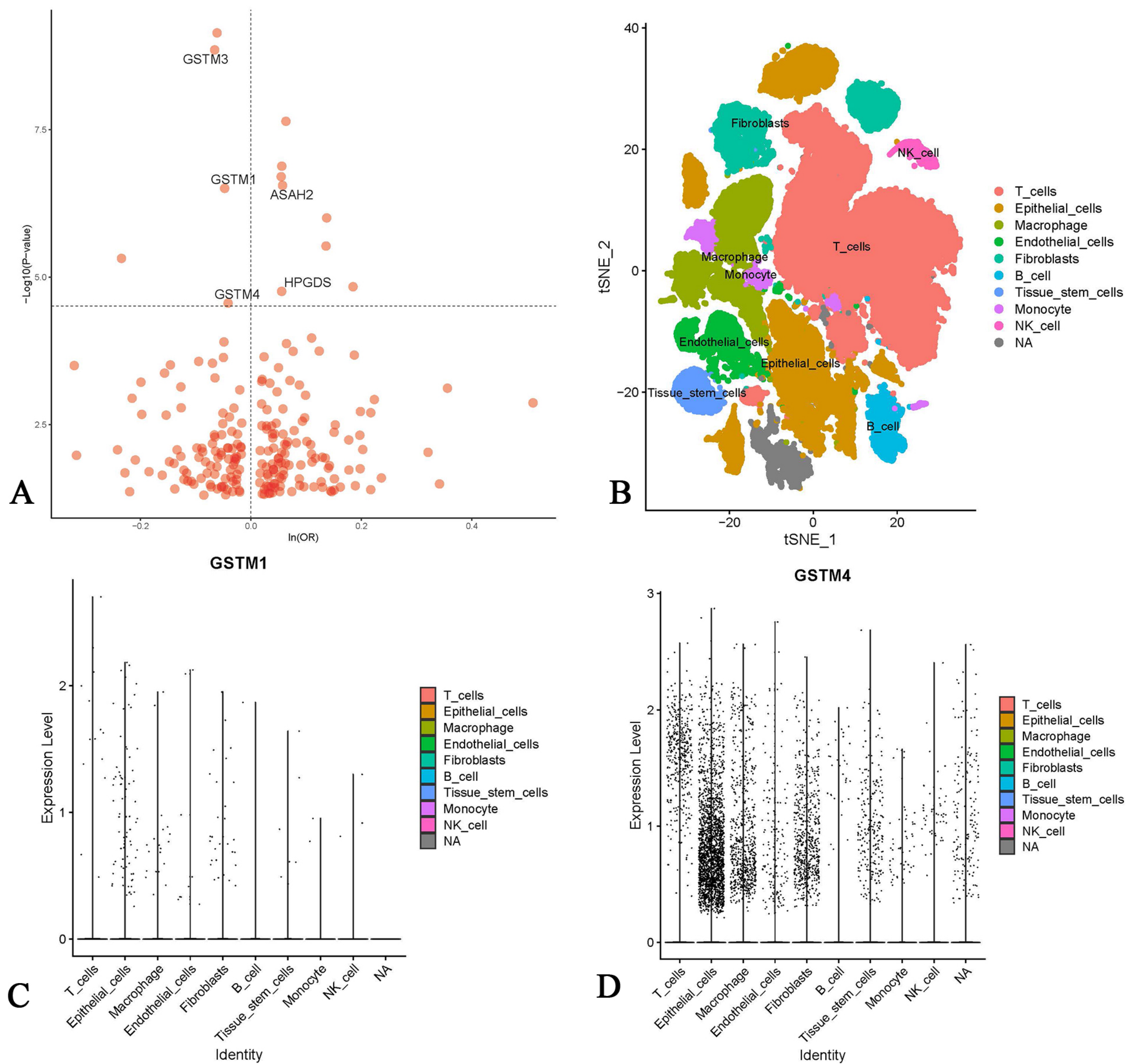


**Fig. 3. Selection of target proteins and single-cell transcriptomic validation.** (A) Volcano plots of the Mendelian Randomization (MR) results for target proteins associated with breast cancer. The horizontal black line represents the Bonferroni correction threshold $-\log_{10}(0.05/1614) = 4.51$. Among GSTM isoforms, GSTM3 demonstrated the strongest protective effect, with all GSTM proteins exhibiting consistent protective directionality. (B) Single-cell RNA sequencing localization analysis of breast cancer (GSE205472), identifying major tumor microenvironment populations, including tumor epithelial cells, fibroblasts, T cells, macrophages, monocytes, B cells, NK cells, tissue stem cells, and endothelial cells. (C,D) Expression patterns of GSTM1 and GSTM4 across identified cell populations.

breast cancer is crucial. Our study integrated proteomic data with bidirectional MR, mediation MR, and Bayesian colocalization to identify therapeutic proteins implicated in breast cancer. Among the identified candidates, PZP, GSTM3, SEMA4D, PENK, ASAH2, TNFSF12, MAX, HPGDS, GSTM1, NID2, SNUPN, GSTM4 and UROD, demonstrated potential therapeutic relevance.

Genetic insights have transformed drug discovery, with genetically validated targets demonstrating greater success in anti-cancer therapy development. Through systematic MR and colocalization investigations, we confirmed multiple proteins associated with breast cancer etiology, supporting their potential as therapeutic candidates based on robust genetic evidence. Notably, we highlighted GSTM, especially GSTM1, as a promising therapeutic intervention. Further analysis is warranted to ensure the reliability of these findings and to explore new therapeutic avenues.

Due to metabolic disturbances and oxidase activation, cancer cells generate elevated levels of reactive oxygen species (ROS). Furthermore, tumor-induced oxidative stress limits immune function and reduces the efficacy of immunotherapies [15]. In addition to protecting against oxidative stress and detoxifying harmful substances, glu-

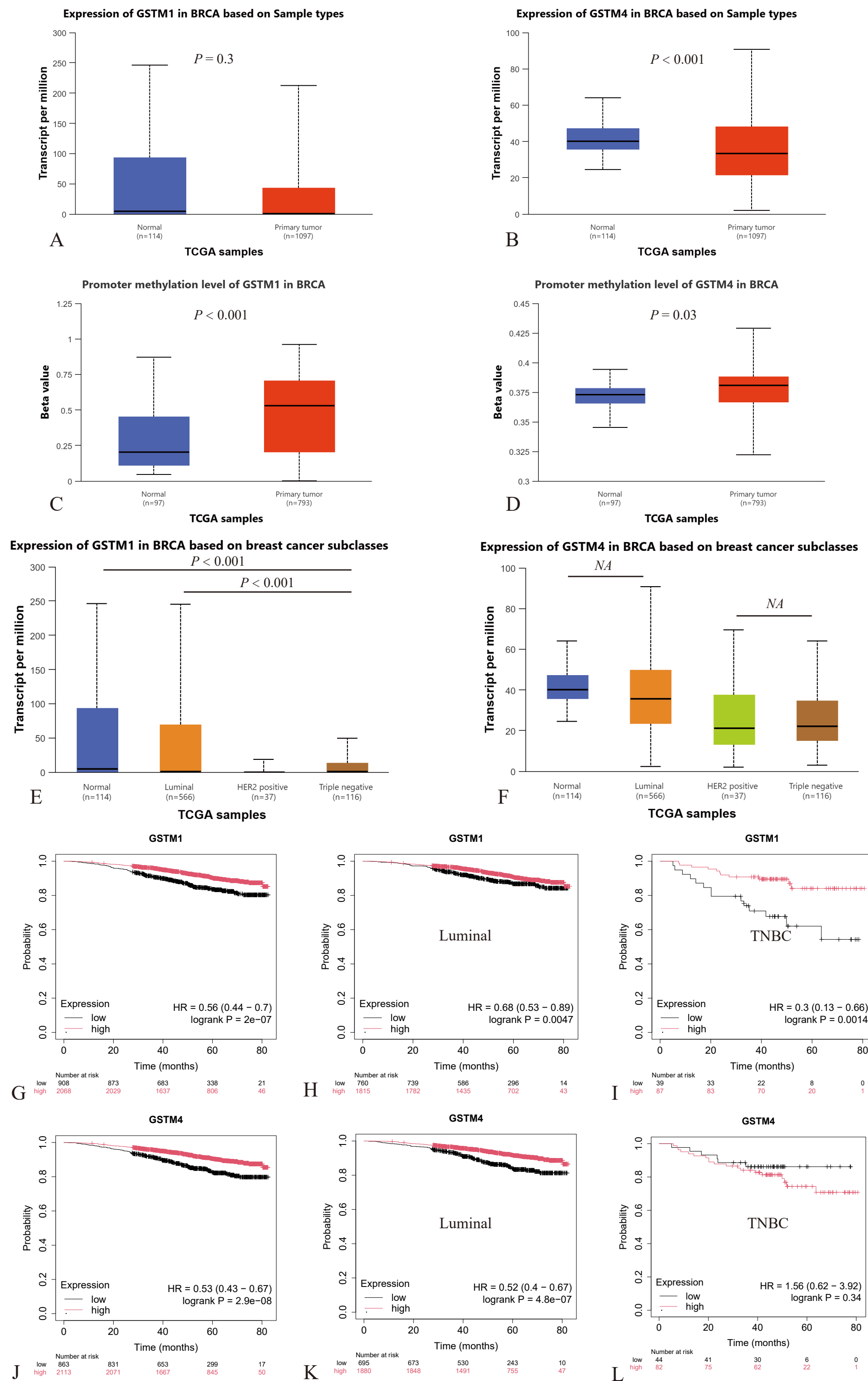


**Fig. 4. GSTM1/4 expression, methylation, and prognostic significance in breast cancer.** (A,B) Comparison of GSTM1 and GSTM4 mRNA expression between breast cancer and normal tissues. (C,D) Promoter methylation levels of GSTM1 and GSTM4, showing hypermethylation in breast cancer ($p < 0.05$). Beta values indicate DNA methylation levels ranging from 0 (unmethylated) to 1 (fully methylated). (E,F) GSTM1 and GSTM4 expression across breast cancer subtypes. In (E), statistically significant group differences are labeled; unlabeled comparisons are nonsignificant. In (F), "NA" indicates nonsignificance; other $p$-values are $< 0.01$. (G–L) Overall survival analysis based on differential expression of GSTM1 and GSTM4 across all breast cancer cases, luminal subtypes, and triple-negative breast cancer (TNBC).

DM DISCOVERY MEDICINE 2001

**Table 3. Candidate drug predicted to target GSTM1 using the Drug Signatures Database (DSigDB).**

| Drug | *p*-value | Adjusted *p*-value | Odds ratio | Combined score | Genes |
|---|---|---|---|---|---|
| Suloctidil | 0.000 | 0.000 | 23.082 | 669.299 | *GSTM1* |
| Arsenenous acid | 0.000 | 0.026 | 3.234 | 31.422 | *GSTM1* |
| Tamibarotene | 0.000 | 0.031 | 4.365 | 39.772 | *GSTM1* |
| Trichostatin | 0.000 | 0.079 | 4.037 | 31.703 | *GSTM1* |
| Decitabine | 0.001 | 0.079 | 2.545 | 19.225 | *GSTM1* |
| Vorinostat | 0.001 | 0.079 | 3.844 | 28.786 | *GSTM1* |
| Apocarotenal | 0.001 | 0.092 | 10.589 | 76.004 | *GSTM1* |
| Midecamycin | 0.001 | 0.133 | 5.411 | 36.162 | *GSTM1* |
| Demecolcine | 0.001 | 0.133 | 3.004 | 19.728 | *GSTM1* |
| Captopril | 0.002 | 0.139 | 6.295 | 40.357 | *GSTM1* |

tathione S-transferases (GSTs) facilitate their conjugation to glutathione [16]. Deficiency of GSTs compromises the ability to efficiently manage oxidative stress, as these enzymes utilize glutathione to alleviate ROS-induced damage [17].

In addition to its traditional roles, GSTM showed antitumor activity that warrants further investigation, especially in breast cancer. Immunotherapy, an essential strategy in breast cancer treatment, faces challenges due to limited efficacy in certain patient populations [18]. Our study illuminated potential interactions between GSTM1/4 and the tumor-associated immune microenvironment (TIME). While TIME serves as a source of therapeutic targets, it can act as both an ally and an adversary to anti-cancer therapies [19].

Single-cell analysis revealed GSTM1/4 enrichment in fibroblasts and epithelial cells, highlighting its potential role in shaping the tumor microenvironment. Mediation MR analysis showed that CD11c on monocytes mediated the causal relationship between GSTM1 and breast cancer. Based on the single-cell analysis results, GSTM1 and GSTM4 did not exhibit significant distinctive expression in monocytes but instead showed low or suppressed expression in monocytes and other immune cells. These findings suggest that the primary function of GSTM1/4 may not depend directly on monocytes but rather on their high expression in epithelial cells and fibroblasts.

CD11c is a specific marker for monocytes and dendritic cells (DCs), typically associated with antigen presentation and immune responses [20]. Research has shown that macrophage aggregation and inflammation in adipose tissue are modulated by nuclear factor-$\kappa$B (NF-$\kappa$B) activation and toll-like receptor-4 (TLR4)-mediated stimulation of macrophages, alongside c-Jun N-terminal kinase (JNK)-related pro-inflammatory pathways in CD11c$^+$ immune cells [21]. CD11c$^+$ monocytes may alter the tumor microenvironment by activating these pro-inflammatory signaling pathways (NF-$\kappa$B, JNK). In combination with the mediation MR results, CD11c appears to mediate the relationship between GSTM1/4 and breast cancer. Collectively, these findings indicate that GSTM1/4 functions may depend not only on the cell types in which they are directly expressed but also indirectly influence breast cancer development through modulation of the TIME. Silencing of GSTM1/GSTM4 in specific cellular subpopulations may represent a novel mechanism underlying breast cancer progression, resistance, and immune evasion.

Although mRNA expression of GSTM1 did not show significant differences, we observed a significantly decreased expression of GSTM4 between breast cancer and normal tissue. Furthermore, we compared promoter methylation levels of these proteins in breast cancer. The results indicated a notably higher level of methylation in GSTM within breast cancer tissues, alongside altered patterns of gene methylation. DNA methylation abnormalities are often observed in cancer cells, including overall hypomethylation and hypermethylation at specific sites, particularly within the CpG islands related to gene regulation, which may help explain these findings [22].

The discordance between GSTM1 protein effects on breast cancer risk and the absence of significant mRNA associations in breast tissue could indicate post-transcriptional regulation mechanisms influencing GSTM1 activity, such as protein stability, post-translational modifications, or tissue-specific translational efficiency, rather than transcriptional control alone. Moreover, survival analysis demonstrated that both GSTM1 and GSTM4 play a protective role in breast cancer, with higher expression correlating with improved outcomes, particularly evident in TNBC based on subgroup analysis. While we utilized TCGA data to explore the prognostic value of GSTM1/4 and methylation, external validation in independent, large-scale cohorts with protein-level and clinical data from real-world settings is needed.

Overall, our findings demonstrate that the expression of GSTM1/4, combined with their significantly elevated methylation levels in breast cancer, provides novel insights into the mechanisms underlying immune evasion. These proteins, potentially regulated by hypermethylation, may experience functional disruption, thereby contributing to breast cancer progression and the development of therapeutic resistance. Additionally, the loss of GSTM1/4 function

may alter the TIME, potentially impairing immune cell activity. These findings underscore the significance of investigating the functional and regulatory networks of GSTM1 and GSTM4 in breast cancer, which could pave the way for the development of personalized therapeutic strategies targeting epigenetic regulation and tumor immunity.

Our study has several limitations. First, the GWAS data utilized in our analysis were derived from large-scale sequencing studies across diverse cohorts, and variations in study protocols may have introduced bias. Second, the data were primarily derived from European populations, limiting the generalizability of our findings to other racial groups. Further studies in non-European populations are necessary to understand the potential clinical applications of these promising targets. Third, a key limitation was the lack of analysis stratified by breast cancer subtypes. Future studies utilizing larger, subtype-specific genomic resources are essential, alongside extensive population-based validation to validate these results.

## Conclusion

A comprehensive investigation identified significant correlations between breast cancer susceptibility and the expression of several proteins, with GSTM, particularly GSTM1, emerging as a critical factor. By establishing a genetic foundation for these potential therapeutic candidates, our work contributes to the development of more precise and effective breast cancer treatments, paving the way for innovative approaches in anti-cancer therapy. As multiple proteins demonstrate suggestive associations, GSTM1/4 are prioritized for further experimental and clinical validation.

## Availability of Data and Materials

The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding authors.

## Author Contributions

QQL and YY contributed to conception and design of the study. QM downloaded the database and jointly participated in the interpretation of the data. YD performed the statistical analysis and wrote the first draft of the manuscript. QQL and YY wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

## Ethics Approval and Consent to Participate

Not applicable.

## Acknowledgment

Not applicable.

## Funding

This research received no external funding.

## Conflict of Interest

The authors declare no conflict of interest.

## Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.24976/Discov.Med.202537199.134.

## References

[1] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA: A Cancer Journal for Clinicians. 2023; 73: 17–48. https://doi.org/10.3322/caac.21763.
[2] Barzaman K, Karami J, Zarei Z, Hosseinzadeh A, Kazemi MH, Moradi-Kalbolandi S, *et al*. Breast cancer: Biology, biomarkers, and treatments. International Immunopharmacology. 2020; 84: 106535. https://doi.org/10.1016/j.intimp.2020.106535.
[3] Guan C, Han Y, Ling Z, Meng X, Zhang B, Dong W, *et al*. Nanomaterials: breaking the bottleneck of breast cancer drug resistance. Frontiers in Immunology. 2024; 15: 1492546. https://doi.org/10.3389/fimmu.2024.1492546.
[4] Birney E. Mendelian Randomization. Cold Spring Harbor Perspectives in Medicine. 2022; 12: a041302. https://doi.org/10.1101/cshperspect.a041302.
[5] Nguyen K, Mitchell BD. A Guide to Understanding Mendelian Randomization Studies. Arthritis Care & Research. 2024; 76: 1451–1460. https://doi.org/10.1002/acr.25400.
[6] Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews. Genetics. 2013; 14: 379–389. https://doi.org/10.1038/nrg3472.
[7] Zheng J, Haberland V, Baird D, Walker V, Haycock PC, Hurle MR, *et al*. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. Nature Genetics. 2020; 52: 1122–1131. https://doi.org/10.1038/s41588-020-0682-6.
[8] Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, *et al*. Large-scale integration of the plasma proteome with genetics and disease. Nature Genetics. 2021; 53: 1712–1721. https://doi.org/10.1038/s41588-021-00978-w.
[9] Orrù V, Steri M, Sidore C, Marongiu M, Serra V, Olla S, *et al*. Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. Nature Genetics. 2020; 52: 1036–1045. https://doi.org/10.1038/s41588-020-0684-4.
[10] Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, *et al*. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nature Communications. 2018; 9: 918. https://doi.org/10.1038/s41467-018-03371-0.
[11] Chandrashekar DS, Karthikeyan SK, Korla PK, Patel H, Shovon AR, Athar M, *et al*. UALCAN: An update to the integrated cancer data analysis platform. Neoplasia. 2022; 25: 18–27. https://doi.org/10.1016/j.neo.2022.01.001.
[12] Győrffy B. Survival analysis across the entire transcriptome

identifies biomarkers with the highest prognostic power in breast cancer. Computational and Structural Biotechnology Journal. 2021; 19: 4101–4109. https://doi.org/10.1016/j.csbj.2021.07.014.

[13] Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, *et al*. DSigDB: drug signatures database for gene set analysis. Bioinformatics. 2015; 31: 3069–3071. https://doi.org/10.1093/bioinformatics/btv313.

[14] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016; 44: W90–W97. https://doi.org/10.1093/nar/gkw377.

[15] O'Reilly A, Zhao W, Wickström S, Arnér ESJ, Kiessling R. Reactive oxygen species: Janus-faced molecules in the era of modern cancer therapy. Journal for Immunotherapy of Cancer. 2024; 12: e009409. https://doi.org/10.1136/jitc-2024-009409.

[16] Hayes JD, Strange RC. Glutathione S-transferase polymorphisms and their biological consequences. Pharmacology. 2000; 61: 154–166. https://doi.org/10.1159/000028396.

[17] Atkinson HJ, Babbitt PC. Glutathione transferases are structural and functional outliers in the thioredoxin fold. Biochemistry. 2009; 48: 11108–11116. https://doi.org/10.1021/bi901180v.

[18] Ye F, Dewanjee S, Li Y, Jha NK, Chen ZS, Kumar A, *et al*. Advancements in clinical aspects of targeted therapy and immunotherapy in breast cancer. Molecular Cancer. 2023; 22: 105. https://doi.org/10.1186/s12943-023-01805-y.

[19] Harris MA, Savas P, Virassamy B, O'Malley MMR, Kay J, Mueller SN, *et al*. Towards targeting the breast cancer immune microenvironment. Nature Reviews. Cancer. 2024; 24: 554–577. https://doi.org/10.1038/s41568-024-00714-6.

[20] Osugi Y, Vuckovic S, Hart DNJ. Myeloid blood CD11c(+) dendritic cells and monocyte-derived dendritic cells differ in their ability to stimulate T lymphocytes. Blood. 2002; 100: 2858–2866. https://doi.org/10.1182/blood.V100.8.2858.

[21] Engin AB. Adipocyte-Macrophage Cross-Talk in Obesity. Advances in Experimental Medicine and Biology. 2017; 960: 327–343. https://doi.org/10.1007/978-3-319-48382-5_14.

[22] Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. Trends in Genetics. 2021; 37: 1012–1027. https://doi.org/10.1016/j.tig.2021.05.002.